
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of Theoretical Biology 269 (2011) 123–130

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

# Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications

Guosen Xie, Zhongxi Mo*

*School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China*

ARTICLE INFO

*Article history:*
Received 15 July 2010
Received in revised form
15 September 2010
Accepted 9 October 2010
Available online 20 October 2010

*Keywords:*
Base content
Projection
Similarity

ABSTRACT

In this article, we introduce three 3D graphical representations of DNA primary sequences, which we call RY-curve, MK-curve and SW-curve, based on three classifications of the DNA bases. The advantages of our representations are that (i) these 3D curves are strictly non-degenerate and there is no loss of information when transferring a DNA sequence to its mathematical representation and (ii) the coordinates of every node on these 3D curves have clear biological implication. Two applications of these 3D curves are presented: (a) a simple formula is derived to calculate the content of the four bases (*A*, *G*, *C* and *T*) from the coordinates of nodes on the curves; and (b) a 12-component characteristic vector is constructed to compare similarity among DNA sequences from different species based on the geometrical centers of the 3D curves. As examples, we examine similarity among the coding sequences of the first exon of beta-globin gene from eleven species and validate similarity of cDNA sequences of beta-globin gene from eight species.



## 1. Introduction

Advances in DNA sequencing technology and DNA databases have greatly facilitated biological research involving DNA sequences. However, it has been acknowledged that information contained in DNA sequences is difficult for humans to comprehend without careful extraction and processing. Many methods have been proposed to characterize DNA sequences, with special efforts given to representing the sequence graphically. Using graphical approaches to study biological problems can provide an intuitive picture or useful insights for helping analyzing complicated relations in these systems, as demonstrated by many previous studies on a series of important biological topics, such as analysis of codon usage (Chou and Zhang, 1992; Zhang and Chou, 1993, 1994), base frequencies in the anti-sense strands (Chou et al., 1996), analysis of DNA and protein sequence (Qi et al., 2007; Wu et al., 2010; Yu et al., 2009), enzyme-catalyzed reactions (Andraos, 2008; Chou, 1980, 1981, 1989; Chou et al., 1979; Chou and Forsen, 1980; Chou and Liu, 1981; Lin and Neet, 1990; Zhou and Deng, 1984), protein folding kinetics and folding rates (Chou, 1990; Chou and Shen, 2009; Shen et al., 2009), inhibition kinetics of processive nucleic acid polymerases and nucleases (Chou et al., 1994), and drug metabolism systems (Chou, 2010). Moreover, graphical methods have also been introduced to deal with some other biological and medical related problems (Diao et al., 2007; Gonzalez-Diaz et al., 2009; Munteanu et al., 2009). Recently, the images of cellular (Wolfram, 1984, 2002) automata were also used to represent biological sequences (Xiao et al., 2005a) for predicting protein structural classes (Xiao et al., 2008) and subcellular location (Xiao et al., 2006b), identifying G-protein-coupled receptor functional classes (Xiao et al., 2009), investigating HBV virus gene missense mutation (Xiao et al., 2005b), HBV viral infections (Xiao et al., 2006a), as well as analyzing SARS-cov (Gao et al., 2006; Wang et al., 2005).

Graphical representation of DNA sequences was first proposed by Hamori and Ruskin (1983). Gates (1986), Nandy (1994) and Leong and Morgenthaler (1995) developed 2D graphical representations of DNA sequences. These methods are straightforward but are accompanied with some loss of information due to overlapping and crossing of the curve representing DNA with itself and degeneracy generated by the circuit. Randic et al. (2003) developed a novel 2D representation method in which there is no loss of information in transferring a DNA sequence to its mathematical representation. More recently, several other 2D representations have been proposed (Wang and Zhang, 2006; Zhang, 2009; Yao et al., 2008; Zhao et al., 2010). As for the 3D graphical representation, Hamori and Ruskin (1983) developed the H-curve. It can uniquely represent a DNA sequence. Based on the classifications of DNA bases, Zhang et al. (Zhang and Zhang, 1994; Zhang, 1997; Zhang et al., 2003) created the Z-curve to represent DNA sequences, in which the four bases (A, G, T and C) are represented by the four vertexes of the regular tetrahedron, as A(1,1,1), T(−1,−1,1), C(−1,1,−1) and G(1,−1,−1). The *Z*-curve is a 3D graphical representation and it has clear biological implication. However, as pointed out by Tang et al. (2010), the *Z*-curve representation has a defeat that it might

* Corresponding author. Tel.: +86 27 68764677.
*E-mail addresses:* xieguosen001@163.com (G. Xie), zhxmo@whu.edu.cn (Z. Mo).

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.10.018

cause a loop in the resulting spatial curve if the frequencies of the four bases present in the sequence are the same. Randic et al. (2000) presented another 3D graphical representation method, but the limitation in forms of crossing and overlapping of the spatial curve representing a DNA sequence still remains. Recently, more other 3D representations were developed by several authors (Li and Wang, 2004; Liao and Wang, 2004; Yu et al., 2009) to overcome the problem of degeneration in graphical representation. These methods, however, do not seem to possess apparent biological meanings.

In this article, we will introduce three novel 3D graphical representations of DNA primary sequences, namely, the RY-curve, the MK-curve and the SW-curve. These curves are derived from three classifications of the four DNA bases A, G, T and C, respectively. It can be proved that the proposed representations are strictly non-degenerate, therefore can avoid potential information loss when transferring a DNA sequence to its representations. Moreover, the coordinates of every node on these 3D curves have clear biological implication. In Section 4, we will present three applications developed based on the proposed representations.

## 2. Construction of RY-curve, MK-curve and SW-curve

The four DNA bases (A, G, T and C) can be classified by the following three ways according to their chemical properties:

(I) Chemical structures of the bases: R (purine)=A, G/Y (pyrimidines)=T,C;
(II) Functional groups of the bases: M (amido)=A, C/K (keto)=G, T.
(III) The strength of the hydrogen bonds between paired bases: S(strong)=G, C/*W*=(weak)A, T.

First consider the R/Y classification. In a 3D space, a point or a vector has three components. We assign the following vectors to the four DNA bases:

$$(1,\ -1,\ 0)\rightarrow \mathrm{A},\ (1,\ 1,\ 0)\rightarrow \mathrm{G},\ (1,\ 0,\ 1)\rightarrow \mathrm{T},\ (1,\ 0,\ -1)\rightarrow \mathrm{C} \tag{1}$$

Notice that we restrict the two vectors representing purine bases R=A,G in the $x$–$y$ plane and two vectors representing pyrimidine bases Y=T,C in the $x$–$z$ plane (see Fig. 1).

Given a DNA sequence with $n$ bases, $S=s_1s_2,\cdots,s_n$, we look at one base at a time. For the $i$-th one ($i=1,2,\ldots,n$), a corresponding point $P_i(x_i,y_i,z_i)$ can be determined in the 3D space as follows:

$$(x_i,y_i,z_i)=\left(\sum_{k=1}^{i}P_k^x,\quad \sum_{k=1}^{i}P_k^y,\quad \sum_{k=1}^{i}P_k^z\right),\quad i=1,2,\ldots,n \tag{2}$$

where $P_k^x$, $P_k^y$ and $P_k^z$ represent the $x$-component, $y$-component and $z$-component of the vector corresponding to $S_k$, respectively. All $n$ bases on the DNA sequence are examined consecutively, and in the end we will obtain $n$ points: $P_1,P_2,\ldots,P_n$ in the 3D space. Then, starting from the original point (0, 0, 0), connecting adjacent points, we will obtain a 3D curve, called as the RY-curve.

As an example, suppose we have a sequence S=ATGGTCTTG. Applying the proposed method, we get ten points corresponding to the nine bases on the sequence (including original point) to be

{(0,0,0), (1,−1,0), (2,−1,1), (3,0,1), (4,1,1), (5,1,2), (6,1,1),
(7,1,2), (8,1,3), (9,2,3)}

Connecting these points sequentially, we obtain the RY-curve (see Fig. 2) for this particular DNA sequence.

Now we consider the M/K classification and the S/W classification of the four bases.

For the M/K classification, we assign the following vectors to the four bases:

$$(1,\ -1,\ 0)\rightarrow \mathrm{A},\ (1,\ 1,\ 0)\rightarrow \mathrm{C},\ (1,\ 0,\ 1)\rightarrow \mathrm{G},\ (1,\ 0,\ -1)\rightarrow \mathrm{T} \tag{3}$$

Here, we restrict two vectors representing the amino bases M=A, C in the $x$–$y$ plane and two vectors representing the keto bases K=G, T in the $x$–$z$ plane. A different way of representing the DNA sequence graphically is thus established. We call the 3D curve generated under this definition the MK-curve.

Similarly, for the S/W classification, we assign the following vectors to the four bases:

$$(1,\ -1,\ 0)\rightarrow \mathrm{A},\ (1,\ 1,\ 0)\rightarrow \mathrm{T},\ (1,\ 0,\ 1)\rightarrow \mathrm{G},\ (1,\ 0,\ -1)\rightarrow \mathrm{C} \tag{4}$$

This time the strong hydrogen bases S=A, T are restricted in the $x$–$y$ plane and the weak hydrogen bases W=G, C are restricted in the $x$–$z$ plane. We then obtain the third 3D graphical representation of the DNA sequence. 3D curves generated under this definition are called the SW-curve.

As an example, in Fig. 3 we plot the RY-curve, MK-curve and SW-curve of human's exon 1 of beta-globin gene in Table 1.

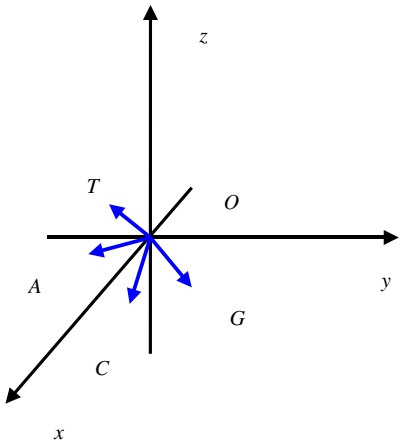


**Fig. 1.** The vectors representing the four bases according to the R/Y classification. Purine bases R=A,G are limited in $x$–$y$ plane and pyrimidine bases Y=T,C are limited in $x$–$z$ plane.

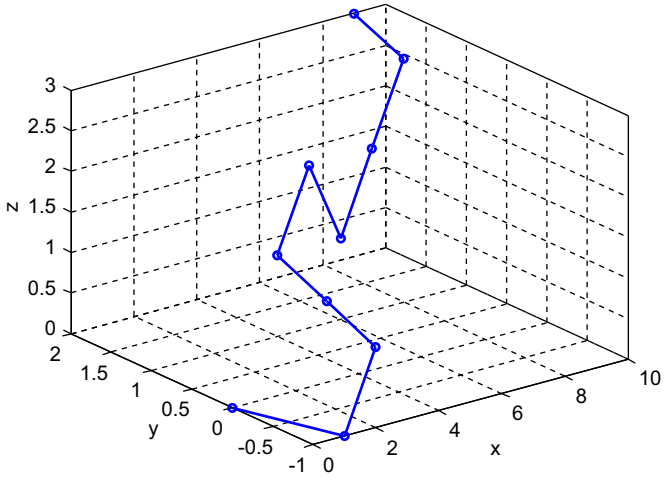


**Fig. 2.** The RY-curve representation of the sequence ATGGTCTTG.

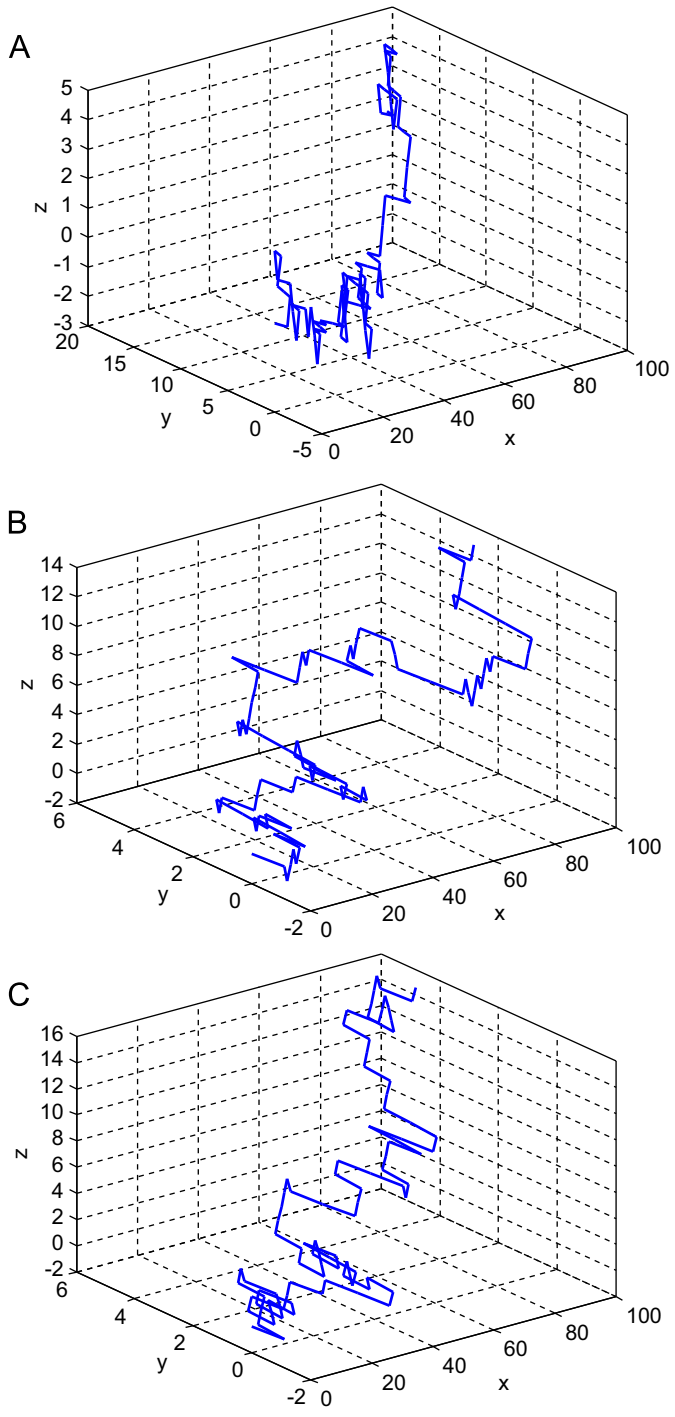


**Fig. 3.** The RY-curve, MK-curve and SW-curve of the coding sequences of the first exon of beta-globin gene of human. (A) The RY-curve of the coding sequences of the first exon of beta-globin gene of human. (B) The MK-curve of the coding sequences of the first exon of beta-globin gene of human. (C) The SW-curve of the coding sequences of the first exon of beta-globin gene of human.

## 3. Properties of RY-curve, MK-curve and SW-curve

In this section, we will prove some properties of RY-curve, MK-curve and SW-curve. We use notations $A_n$, $G_n$, $T_n$ and $C_n$ to denote the content of bases A, T, G and C, respectively, in a DNA sequence: $S=s_1s_2\cdots s_n$, $s_i\in\{A,T,G,C\}$.

**Property 3.1.** *There is no circuit and degeneracy in RY-curve, MK-curve and SW-curve.*

**Proof.** We can prove this property by contradiction. First consider the RY-curve. Suppose that there are one or more circuits in a RY-curve. Then there exists at least one point in the 3D space at which the curve crosses itself. That is, two points on the curve, say $P_i$ and $P_j$, $i\neq j$, have exactly the same coordinates $(x_i,y_i,z_i)=(x_j,y_j,z_j)$. So we must have $x_i=x_j$. According to the Assignment (1) and Eq. (2), we have $x_i=\sum_{k=1}^{i}p_k^x=i$ and $x_j=\sum_{k=1}^{j}P_k^x=j$. This implies $i=j$. However, this contradicts the supposition that $i\neq j$. Therefore, there is no circuit and degeneracy in RY-curve.

Similarly, we can show that there is also no circuit and degeneracy in MK-curve and SW-curve. □

**Property 3.2.** *There exists an one-to-one correspondence between a DNA sequence and a RY-curve (MK-curve or SW-curve) and no loss of information is resulted.*

**Proof.** First consider RY-curve. From the previous proof, we know that, for a given DNA sequences $S=s_1s_2\cdots s_n$, there exists one unique RY-curve. □

Conversely, suppose that RY-curve of a DNA sequence is given; it then follows immediately that the coordinates of all $n$ nodes on the RY-curve, $(x_i,y_i,z_i)$, $i=1,2,\ldots,n$ are given. Let $(x_0,y_0,z_0)=(0,0,0)$. According to Eq. (2), bases $s_i$ corresponding to the node $P(x_i,y_i,z_i)$ on the RY-curve can be calculated using the following formula:

$$s_i=\begin{cases}\mathrm{A},(x_i,y_i,z_i)-(x_{i-1},y_{i-1},z_{i-1})=(1,-1,0)\\ \mathrm{G},(x_i,y_i,z_i)-(x_{i-1},y_{i-1},z_{i-1})=(1,1,0)\\ \mathrm{T},(x_i,y_i,z_i)-(x_{i-1},y_{i-1},z_{i-1})=(1,0,1)\\ \mathrm{C},(x_i,y_i,z_i)-(x_{i-1},y_{i-1},z_{i-1})=(1,0,-1)\end{cases},\quad i=1,2,\ldots,n \tag{5}$$

Formula (5) consists of the followings set of equations:

$$\begin{cases}(x_1y_1z_1)-(x_0y_0z_0)=\alpha_1\\ (x_2y_2z_2)-(x_1y_1z_1)=\alpha_2\\ \vdots\\ (x_ny_nz_n)-(x_{n-1}y_{n-1}z_{n-1})=\alpha_n\end{cases} \tag{6}$$

where $a_i\in\{(1,-1,0),(1,1,0),(1,0,1),(1,0,-1)\}$ and $i=1,2,\ldots,n$ is known. Note $(x_0,y_0,z_0)=(0,0,0)$. Regarding $(x_1,y_1,z_1))\cdots(x_n,y_n,z_n)$ as unknown, we obtain the coefficient matrix of the Eq. (6) to be

$$A=\begin{bmatrix}1&0&0&\cdots&0\\ -1&1&0&\cdots&0\\ 0&-1&1&\cdots&0\\ \vdots&\vdots&\vdots&\vdots&\vdots\\ 0&0&0&\cdots&1\end{bmatrix}$$

The determinant $|A|=1\neq 0$, therefore for the given RY-curve, Eqs. (6) have a unique solution. This implies that one RY-curve uniquely determines one correspondent DNA sequence. Hence, the correspondence between DNA sequences and RY-curves is one-to-one and there is no loss of information.

Similarly, we can prove that Property 3.2 holds for MK-curve and SW-curve as well.

**Property 3.3.** *The x-component of the vector corresponding to the node $P(x_n,y_n,z_n)$ of the RY-curve (MK-curve or SW-curve), $x_n$, is just the length of the DNA sequence $S=s_1s_2\cdots s_n$, we have*

$A_n+G_n+T_n+C_n=x_n$

**Proof.** The proof follows immediately from assignment (1) and Eq. (2). □

**Property 3.4.**

(i) *For the RY-curve, its projections (2D curve) onto the x–y plane and the x–z plane denote the distributions of purine bases (A,G) and pyrimidine bases (T,C) along the sequence $S=s_1s_2,\ldots,s_n$,*

**Table 1**
The coding sequences of the first exon of beta-globin gene of 11 different species.

| Species | Coding sequences |
|---|---|
| Human | ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG |
| Goat | ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG |
| Gallus | ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG |
| Mouse | ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG |
| Rat | ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG |
| Chimpanzee | ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG |
| Bovine | ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG |
| Gorilla | ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG |
| Opossum | ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG |
| Lemur | ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG |
| Rabbit | ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC |

*respectively, and we have*

$G_n - A_n = y_n, \quad T_n - C_n = z_n$

(ii) *For the MK-curve, its projections (2D curve) onto the x–y plane and the x–z plane denotes the distributions of bases of amino group (A,C) and keto group (G,T) along the sequence* $S = s_1 s_2, \ldots, s_n$, *respectively, and we have*

$C_n - A_n = y_n, \quad G_n - T_n = z_n$

(iii) *For the SW-curve, its projections (2D curve) onto the x–y plane and the x–z plane denotes the distributions of bases of weak H-bonds (A,T) and strong H-bonds (G,C) along the sequence* $S = s_1 s_2, \ldots, s_n$, *respectively, and we have*

$T_n - A_n = P_n^y, \quad G_n - C_n = P_n^z$

**Proof.** First we prove (i). From assignment (1), we know that the vectors representing bases, G and A, are symmetrical about x axis on the x-y plane, and the vectors representing bases, T and C, also are symmetrical about x axis on the x-z plane. So, according to Equation (2), we have

$G_n - A_n = y_n, \quad \text{and} \quad T_n - C_n = z_n$

The projection of RY-curve onto the $x$–$y$ plane is a 2D curve with nodes: $\{(x_k, y_k), k = 1,2,\ldots,n\}$ and the projection of RY-curve onto the $x$–$z$ plane is a 2D curve with nodes:$\{(x_k, z_k), k = 1,2,\ldots,n\}$. So, the projections of RY-curve onto the $x$–$y$ plane and the $x$–$z$ plane display the distributions of purine bases (A,G) and pyrimidine bases (T,C) along the sequence, respectively. □

Similarly, we can prove (ii) and (iii).

Fig. 4 shows the projections of the RY-curve of human's exon 1 of beta-globin genes onto the $x$–$y$ plane and the $x$–$z$ plane.

From Fig. 4, we can see that $G_n > A_n$, $T_n > C_n$, and the changing trend of the content of bases A,G and bases T,C along the sequence is also directly observable.

**Property 3.5.** *Let* $(x_n, y_n, z_n)$ *denote the coordinates of the terminal point on an RY-curve, then the following relationships are true:*

(i) *if* $y_n > 0, z_n > 0$, *then* $T_n + G_n > A_n + C_n$.
(ii) *if* $y_n < 0, z_n < 0$, *then* $T_n + G_n < A_n + C_n$.
(iii) *if* $y_n > 0$, $z_n < 0$, *then* $G_n + C_n > A_n + T_n$
(iv) *if* $y_n < 0$, $z_n > 0$, *then* $G_n + C_n < A_n + T_n$
(v) *if* $y_n = 0$, $z_n = 0$, *then* $T_n + G_n = A_n + C_n$ *and* $T_n = C_n$, $G_n = A_n$.

**Proof.** The above results follow from assignment (1) and Eq. (2) directly.

For MK-curve and SW-curve, there are analogous properties as well. □

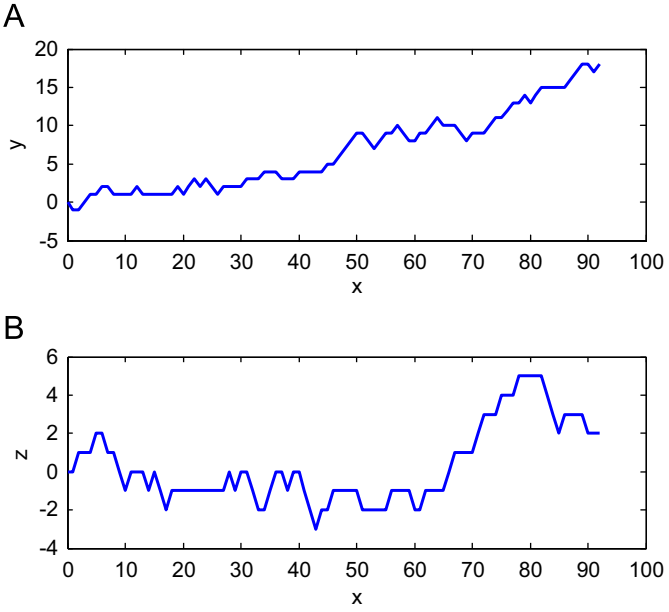


**Fig. 4.** The 2D projection of the RY-curve of the first exon of beta-globin gene of human in $x$–$y$ plane and $x$–$z$ plane. (A) The 2D projection of the RY-curve of the first exon of beta-globin gene of human in $x$–$y$ plane, which denotes the distribution of purine bases (A,G) along the coding sequence. (B) The 2D projection of the RY-curve of the first exon of beta-globin gene of human in $x$–$z$ plane, which denotes the distribution of pyrimidine bases (T,C) along the coding sequence.

## 4. The applications of RY-curve, MK-curve and SW-curve

In this section, we will present two applications of RY-curve, MK-curve and SW-curve.

### 4.1. Calculation of the base content of a DNA sequence

Based on Properties 3.3 and 3.4, we can obtain three equations for each curve.

For RY-curve, we have

$$\begin{cases} A_n + G_n + T_n + C_n = x_n^{(RY)} \\ G_n - A_n = y_n^{(RY)} \\ T_n - C_n = z_n^{(RY)} \end{cases} \tag{7}$$

For MK-curve, we have

$$\begin{cases} A_n + G_n + T_n + C_n = x_n^{(MK)} \\ C_n - A_n = y_n^{(MK)} \\ G_n - T_n = z_n^{(MK)} \end{cases} \tag{8}$$

And for SW-curve, we have

$$\begin{cases} A_n+G_n+T_n+C_n=x_n^{(SW)} \\ T_n-A_n=y_n^{(SW)} \\ G_n-C_n=z_n^{(SW)} \end{cases} \quad (9)$$

Without loss of generality, we select the following four independent equations from (7)–(9):

$$\begin{cases} A_n+G_n+T_n+C_n=x_n^{(RY)} \\ -A_n+G_n=y_n^{(RY)} \\ G_n-T_n=z_n^{(MK)} \\ T_n-C_n=z_n^{(RY)} \end{cases} \quad (10)$$

Notice that since the coefficient matrix of Eq. (10) is nonsingular, there exists one unique solution.

The solution of Eq. (10) can be obtained recursively as follows:

$$\begin{cases} C_n=\frac{1}{4}\left(x_n^{(RY)}+y_n^{(RY)}\right)-\frac{3}{4}z_n^{(RY)}-\frac{1}{2}z_n^{(MK)} \\ T_n=\frac{1}{3}\left(x_n^{(RY)}+y_n^{(RY)}\right)-\frac{2}{3}z_n^{(MK)}-\frac{1}{3}C_n \\ G_n=\frac{1}{2}\left(x_n^{(RY)}+y_n^{(RY)}\right)-\frac{1}{2}T_n-\frac{1}{2}C_n \\ A_n=x_n^{(RY)}-G_n-T_n-C_n \end{cases} \quad (11)$$

For example, for the complete coding sequences of beta-globin genes of human, from its RY-curve and MK-curve, we obtain

$$x_n^{(RY)}=92,\quad y_n^{(RY)}=18,\quad z_n^{(RY)}=2,\quad z_n^{(MK)}=14.$$

Substituting these values into formula (11), we get

$$(A_n,\ G_n,\ T_n,\ C_n)=(17,\ 35,\ 21,\ 19).$$

Similarly, using formula (11), we can calculate the base content of DNA sequences for the eleven species presented in Table 1 (see Table 2).

### 4.2. Similarity analysis based on the RY-curve, MK-curve and SW-curve

Comparing similarities among different DNA sequences is one of the essential motivations of graphical representation. In order to do this, Randic et al. (2000) proposed E matrix, M/M matrix, L/L matrix and $L^k/L^k$ matrix. They used matrix eigenvalues as the sequence descriptors to make comparisons among DNA sequences. This method has been proved to be useful and used by many authors (Randic et al., 2000, 2003; Wang and Zhang, 2006;Li and Wang, 2004). However, when DNA sequence is very long, these matrices can become very large, and the calculation could become very complicated. Yao et al. (2005) used the coordinates of the geometric center of graph as the sequence descriptors to do similarity analysis among different DNA sequences. The method is simple as far as calculation is concerned. However, it does have a potentially serious drawback: when the two DNA sequences under comparison contain the same proportions of the bases A, G, T and C, they may have the same geometric center, although they can be completely different. To overcome this unfavorable drawback, we develop a new method for comparing similarity between two DNA sequences based on our proposed RY-curve, MK-curve and SW-curve. A twelve-component vector that serves as a sequence descriptor is constructed based on the geometrical centers of the representing curves.

**Table 2**
The base contents of the 11 coding sequences of Table 1.

| Species | A | G | T | C | Total |
|---|---|---|---|---|---|
| Human | 17 | 35 | 21 | 19 | 92 |
| Goat | 17 | 35 | 17 | 17 | 86 |
| Gallus | 19 | 34 | 15 | 24 | 92 |
| Mouse | 17 | 34 | 23 | 20 | 94 |
| Rat | 20 | 33 | 21 | 18 | 92 |
| Chimpanzee | 20 | 41 | 24 | 20 | 105 |
| Bovine | 17 | 35 | 18 | 16 | 86 |
| Gorilla | 17 | 37 | 20 | 19 | 93 |
| Opossum | 21 | 29 | 22 | 20 | 92 |
| Lemur | 19 | 35 | 23 | 15 | 92 |
| Rabbit | 17 | 37 | 20 | 16 | 90 |

#### 4.2.1. Construction of the 12-component sequence descriptor

In Section 2, we have constructed the RY-curve of representing a DNA sequence restricting purine bases R=A,G in the $x$–$y$ plane and pyrimidine bases Y=T,C in the $x$–$z$ plane. Conditional on this assumption, there exist four possible ways of assigning the four vectors to the four bases (A, G, T and C):

$$\begin{cases} (1,-1,0)\to A,\ (1,1,0)\to G,\ (1,0,1)\to T,\ (1,0,-1)\to C \\ (1,-1,0)\to G,\ (1,1,0)\to A,\ (1,0,1)\to T,\ (1,0,-1)\to C \\ (1,-1,0)\to A,\ (1,1,0)\to G,\ (1,0,1)\to C,\ (1,0,-1)\to T \\ (1,-1,0)\to G,\ (1,1,0)\to A,\ (1,0,1)\to C,\ (1,0,-1)\to T \end{cases} \quad (12)$$

Thus, from assignment (12) and Eq. (2), we could have four different kinds of RY-curve, denoted as RY-curve11, RY-curve12, RY-curve13 and RY-curve14. Note these curves are listed in the same order as they appear in Eq. (2).

Analogously, for MK-curve, we can also form four kinds of MK-curve, denoted by MK-curve21, MK-curve22, MK-curve23 and MK-curve24. For SW-curve, we can also obtain four kinds of SW-curve, denoted by SW-curve31, SW-curve32, SW-curve33 and SW-curve34. Therefore, we can have a total of twelve 3D curves representing a DNA sequence.

For a given sequence with length $n$, we have a set of points $(x_i, y_i, z_i)$, $i=1, 2, 3,\ldots, n$, from the graphical representation of the sequence. The coordinates of the geometrical center of all the points, denoted by $x_0$, $y_0$, and $z_0$, can be calculated as follows (Yao et al., 2005):

$$(x_0,y_0,z_0)=\left(\frac{1}{n}\sum_{i=1}^{n}x_i,\frac{1}{n}\sum_{i=1}^{n}y_i,\frac{1}{n}\sum_{i=1}^{n}z_i\right) \quad (13)$$

Next, we calculate the following index by (13):

$$I=\sqrt[3]{\sum_{i=1}^{n}(x_i-x_0)(y_i-y_0)(z_i-z_0)/n} \quad (14)$$

Using formula (14), we calculate an index vector based on all above twelve 3D curves, denoted by

$$\vec{d}=(I_{11},I_{12},I_{13},I_{14},I_{21},I_{22},I_{23},I_{24},I_{31},I_{32},I_{33},I_{34}) \quad (15)$$

Here, we use the first subscript to denote the particular curve (RY, MK, SW) and use the second subscript to denote the four different ways concerning how the vectors are assigned. The 12-component vector (15) can be used as the sequence descriptors. To ease notational exposition, we rewrite the 12-component vector (15) as follows:

$$\vec{d}=(I_1,I_2,I_3,I_4,I_5,I_6,I_7,I_8,I_9,I_{10},I_{11},I_{12}) \quad (15')$$

#### 4.2.2. Similarity analysis of the coding sequences of beta-globin gene among different species

Comparison based on sequence descriptors is one method, which has been routinely used in similarity analysis. Here, we use the 12-component vector (15) as the index for comparing different DNA sequences.

Suppose that for species $i$ and $j$, their 12-component vectors are

$$\vec{d}_i=(I_{i1},I_{i2},I_{i3},I_{i4},I_{i5},I_{i6},I_{i7},I_{i8},I_{i9},I_{i10},I_{i11},I_{i12})$$

and

$$\vec{d}_j = (I_{j1}, I_{j2}, I_{j3}, I_{j4}, I_{j5}, I_{j6}, I_{j7}, I_{j8}, I_{j9}, I_{j10}, I_{j11}, I_{j12})$$

We introduce two measures to quantify similarity between the two species. They are the Euclidean distance $d_{ij}$ and the correlation angle $\theta_{ij}$:

$$d_{ij} = \sqrt{(I_{i1}-I_{j1})^2 + \cdots + (I_{i12}-I_{j12})^2}$$

$$\cos\theta_{ij} = \sum_{k=1}^{12} I_{ik}I_{jk} \Big/ \left( \sqrt{\sum_{k=1}^{12} I_{ik}{}^2} \sqrt{\sum_{k=1}^{12} I_{jk}{}^2} \right)$$

The smaller the $d_{ij}$ and $\theta_{ij}$ are, the more the similar species $i$ and $j$ are.

Calculating $d_{ij}$ and $\theta_{ij}$ for all eleven species presented in Table 1, we obtain two similarity matrices: $M1$ and $M2$, where $M1=(d_{ij})_{11\times 11}$ and $M2=(\theta_{ij})_{11\times 11}$. To combine information from these two matrices together, we compute a weighted sum: $M(a)=aM1+(1-a)M2$, $(0\le a\le 1)$, as the overall similarity matrix of the eleven species. Setting $a=1/2$, we compute the overall similarity matrix $M(1/2)$ for the eleven species and list the result in Table 3.

It can be observed in Table 3 that the following pairs of species have significantly smaller similarity scores: human–chimpanzee, human–gorilla, gorilla–chimpanzee and bovine–gorilla. Gallus (the only nonmammalian species) is greatly dissimilar with the other species except for rat, because all other entries involving gallus are relatively large (see the fourth row of Table 3). It is also observed that opossum has larger similarity scores with other species (see the tenth column of Table 3). As presented in Table 3, human–goat, human–bovine, goat–bovine, goat–gorilla, chimpanzee–bovine and chimpanzee–gorilla have smaller entries, so they are only moderately similar.

To compare our results with others, we list the currently published results on comparing the similarity of human and other several species in Table 4. As one can see, there is only limited variation among these different methods, therefore these methods are in overall agreement.

#### 4.2.3. Similarity analysis of cDNA sequences of beta-globin gene among different species

Based on the method proposed in Section 4.2.1., we compare similarities among cDNA sequences of beta-globin gene of eight species in Table 5. The results are listed in Table 6.

It can be observed in Table 6 that the following pairs of species have significantly smaller similarity scores: human–chimpanzee, rat–mouse and mouse lemur–bushbaby. In fact, the eight species chosen here are four pairs of close relatives in their evolution, namely human–chimpanzee, rat–mouse, mouse lemur –bushbaby and tetraodon–fugu. However, we notice that the similarity score of tetraodon–fugu is the smallest in the seventh column of Table 6, but it is much bigger than the other three close relative entries. This problem remains to be further studied.

## 5. Conclusion

In this paper, we propose three graphical representations, namely RY-curve, MK-curve and SW-curve, to represent the DNA sequence in a 3D space. We prove that the 3D curves are strictly non-degenerate and there is no loss of information in transferring the DNA sequence to the proposed curves. Compared with other graphical representations, the main advantage of our method is

**Table 5**
The cDNA sequences of beta-globin gene of 8 species.

| Species | Release date | UCSC version | Length (bp) |
|---|---|---|---|
| Human | Feb. 2009 | hg19/GRCh37 | 444 |
| Chimpanzee | Mar. 2006 | panTro2 | 444 |
| Rat | Nov. 2004 | rn4 | 444 |
| Mouse | July 2007 | mm9 | 444 |
| Tetraodon | Mar. 2007 | tetNig2 | 448 |
| Fugu | Oct. 2004 | fr2 | 444 |
| Mouse lemur | Jun. 2003 | micMur1 | 443 |
| Bushbaby | Dec. 2006 | otoGar1 | 444 |

**Table 3**
The similarity matrix of the 11 coding sequences of Table 1: $M(1/2)$.

| Species | Human | Goat | Gallus | Mouse | Rat | Chimpanzee | Bovine | Gorilla | Opossum | Lemur | Rabbit |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Human | 0 | 0.0789 | 1.1475 | 1.489 | 1.0999 | 0.0062 | 0.0735 | 0.0424 | 0.6468 | 0.5246 | 0.3757 |
| Goat | | 0 | 1.2189 | 1.5666 | 1.1719 | 0.0736 | 0.0057 | 0.0367 | 0.7203 | 0.4509 | 0.3034 |
| Gallus | | | 0 | 0.3509 | 0.0478 | 1.1529 | 1.2142 | 1.1849 | 0.6076 | 1.4876 | 1.5201 |
| Mouse | | | | 0 | 0.397 | 1.4951 | 1.5617 | 1.5305 | 0.8499 | 1.1405 | 1.2874 |
| Rat | | | | | 0 | 1.1054 | 1.1671 | 1.1376 | 0.5699 | 1.5337 | 1.4727 |
| Chimpanzee | | | | | | 0 | 0.0681 | 0.0371 | 0.6528 | 0.52 | 0.371 |
| Bovine | | | | | | | 0 | 0.0314 | 0.7159 | 0.4563 | 0.3086 |
| Gorilla | | | | | | | | 0 | 0.6872 | 0.486 | 0.3379 |
| Opossum | | | | | | | | | 0 | 1.0483 | 0.9319 |
| Lemur | | | | | | | | | | 0 | 0.1497 |
| Rabbit | | | | | | | | | | | 0 |

**Table 4**
The degree of similarity of the coding sequences of several species with the coding sequences of human.

| Species | Chimpanzee | Gorilla | Gallus | Opossum | Bovine | Goat |
|---|---|---|---|---|---|---|
| Our work, Table 3 | 0.0062 | 0.0424 | 1.1475 | 0.6468 | 0.0735 | 0.0789 |
| Liao and Ding (2006), Table 5 | 0.022893 | 0.025960 | 0.0106123 | 0.095765 | 0.048664 | 0.052039 |
| Liu et al. (2006), Table 5 | 0.0145 | 0.0079 | 0.2417 | 0.2815 | 0.0750 | 0.1078 |
| Yao et al. (2008), Table 10 | 0.00449 | 0.00478 | 0.02916 | 0.02999 | 0.01359 | 0.01633 |
| Zhang (2009), Table 1 | 0.9572 | 0.2633 | 1.1559 | 1.1863 | 0.3606 | 0.4769 |
| Tang et al. (2010), Table 3 | 0.0399 | 0.0441 | 0.1766 | 0.1598 | 0.0799 | 0.0869 |
| Tang et al. (2010), Table 4 | 0.0379 | 0.0423 | 0.1781 | 0.1598 | 0.0796 | 0.0855 |

**Table 6**
The similarity matrix of the cDNA of 8 species in Table 5: $M(1/2)$.

| Species | Human | Chimpanzee | Rat | Mouse | Tetraodon | Fugu | Mouse lemur | Bushbaby |
|---|---|---|---|---|---|---|---|---|
| Human | 0 | 0.013149 | 0.36733 | 0.44743 | 0.60224 | 1.4339 | 0.37625 | 0.42344 |
| Chimpanzee | | 0 | 0.38038 | 0.46053 | 0.61479 | 1.4465 | 0.38932 | 0.43653 |
| Rat | | | 0 | 0.083386 | 0.27248 | 1.0744 | 0.009278 | 0.057991 |
| Mouse | | | | 0 | 0.19741 | 0.99561 | 0.074264 | 0.025603 |
| Tetraodon | | | | | 0 | 0.83818 | 0.26465 | 0.22101 |
| Fugu | | | | | | 0 | 1.066 | 1.0195 |
| Mouse lemur | | | | | | | 0 | 0.048804 |
| Bushbaby | | | | | | | | 0 |

that the 2D projection of RY-curve, MK-curve and SW-curve onto the $x$–$y$ plane and the $x$–$z$ plane has clear biological implication. For example, the 2D projection of RY-curve onto the $x$–$y$ plane denotes the changing trend of the content of A, G (see Fig. 4). The three components of the terminal node of these 3D curves algebraically relate to the content of the bases: A, G, T and C (see Properties 3.3 and 3.4). Therefore, more information is retained by our method compared to other available methods. As the application of the graphical representation, we derive a simple formula to recover the content of the four kinds of bases (A, G, C and T) in a DNA sequence from the proposed curves. The sequence descriptors of 12-component vectors we have constructed enabled us to conduct similarity analysis among the coding sequences of first exon of beta-globin gene of 11 species. Our results are in overall agreement with the results reported in the article (Zhang, 2009; Yao et al., 2008; Tang et al., 2010; Liao and Ding, 2006; Liu et al., 2006) (see Table 4). We also have a good validation of similarities of cDNA sequences of the related ones by our method. Computation involved in implementing the proposed methods is fairly straightforward.

## References


Andraos, J., 2008. Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. Can. J. Chem. 86, 342–357.

Chou, K.C., Zhang, C.T., 1992. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. AIDS Res. Human Retrovir. 8, 1967–1976.

Chou, K.C., Zhang, C.T., Elrod, D.W., 1996. Do antisense proteins exist? J. Protein Chem. 15 59–61.

Chou, K.C., 1980. A new schematic method in enzyme kinetics. Eur. J. Biochem. 113, 195–198.

Chou, K.C., 1981. Two new schematic rules for rate laws of enzyme-catalyzed reactions. J. Theor. Biol. 89, 581–592.

Chou, K.C., 1989. Graphic rules in steady and non-steady enzyme kinetics. J. Biol. Chem. 264, 12074–12079.

Chou, K.C., Jiang, S.P., Liu, W.M., Fee, C.H., 1979. Graph theory of enzyme kinetics: 1. Steady-state reaction system. Sci. Sin. 22, 341–358.

Chou, K.C., Forsen, S., 1980. Graphical rules for enzyme-catalyzed rate laws. Biochem. J. 187, 829–835.

Chou, K.C., Liu, W.M., 1981. Graphical rules for non-steady state enzyme kinetics. J. Theor. Biol. 91, 637–654.

Chou, K.C., Shen, H.B., 2009. FoldRate: a web-server for predicting protein folding rates from primary sequence. The Open Bioinform. J. 3, 31–50.

Chou, K.C., 1990. Review: applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. Biophysical Chemistry 35, 1–24.

Chou, K.C., Kezdy, F.J., Reusser, F., 1994. Review: steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. Anal. Biochem. 221, 217–230.

Chou, K.C., 2010. Graphic rule for drug metabolism systems. Curr. Drug Metabol. 11, 369–378.

Diao, Y., Li, M., Feng, Z., Yin, J., Pan, Y., 2007. The community structure of human cellular signaling network. J. Theor. Biol. 247, 608–615.

Gao, L., Ding, Y.S., Dai, H., Shao, S.H., Huang, Z.D., Chou, K.C., 2006. A novel fingerprint map for detecting SARS-CoV. J. Pharmaceut. Biomed. Anal. 41, 246–250.

Gate, M., 1986. A simple way to look at DNA. J. Theor. Biol., 319–328.

Gonzalez-Diaz, H., Perez-Montoto, L.G., Duardo-Sanchez, A., Paniagua, E., Vazquez-Prieto, S., Vilas, R., Dea-Ayuela, M.A., Bolas-Fernandez, F., Munteanu, C.R., Dorado, J., Costas, J., Ubeira, F.M., 2009. Generalized lattice graphs for 2D-visualization of biological information. J. Theor. Biol. 261, 136–147.

Hamori, E., Ruskin, J., 1983. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. J. Biol. Chem. 25, 1318–1327.

Leong, P.M., Morgenthaler, S., 1995. Random walk and gap plots of DNA sequences. Comput. Appl. Biosci. 11, 503–507.

Li, C., Wang, J., 2004. On a 3-D representation of DNA primary sequences. Comb. Chem. High Throughput Screen. 7, 23–27.

Liao, B., Ding, K.Q., 2006. A 3D graphical representation of DNA sequences and its application. Theor. Comput. Sci. 358, 56–64.

Liao, B., Wang, T.M., 2004. 3D graphical representation of DNA sequences and their numerical characterization. J. Mol. Struct. 681, 209–212.

Lin, S.X., Neet, K.E., 1990. Demonstration of a slow conformational change in liver glucokinase by fluorescence spectroscopy. J. Biol. Chem. 265, 9670–9675.

Liu, X.Q., Dai, Q., Xiu, Z.L., Wang, T.M., 2006. PNN-curve: a new 2D graphical representation of DNA sequences and its application. J. Theor. Biol. 243, 555–561.

Munteanu, C.R., Magalhaes, A.L., Uriarte, E., Gonzalez-Diaz, H., 2009. Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. J. Theor. Biol. 257, 303–311.

Nandy, A., 1994. A new graphical representation and analysis of DNA sequence structure. I: methodology and application to globin genes. Curr. Sci. 66, 309–314.

Qi, X.Q., Wen, J., Qi, Z.H., 2007. New 3D graphical representation of DNA sequence based on dual nucleotides. J. Theror. Biol. 249, 681–690.

Randic, M., Vracko, M., Lers, N., Plavsic, O., 2000. On 3-D graphical representation of DNA primary sequence and their numerical characterization. J. Chem. Inform.-Comput. Sci. 40, 1235–1244.

Randic, M., Vracko, M., Lers, N., Plavsic, O., 2003. Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem. Phys. Lett. 368, 1–6.

Shen, H.B., Song, J.N., Chou, K.C., 2009. Prediction of protein folding rates from primary sequence by fusing multiple sequential features. J. Biomed. Sci. Eng. (JBiSE) 2, 136–143.

Tang, X.C., Zhou, P.P., Qiu, W.Y., 2010. On the similarity/dissimilarity of DNA sequences based on 4D graphical representation. Chin. Sci. Bull. 55, 701–704.

Wang, M., Yao, J.S., Huang, Z.D., Xu, Z.J., Liu, G.P., Zhao, H.Y., Wang, X.Y., Yang, J., Zhu, Y.S., Chou, K.C., 2005. A new nucleotide-composition based fingerprint of SARS-CoV with visualization analysis. Med. Chem. 1, 39–47.

Wang, J., Zhang, Y., 2006. Characterization and similarity analysis of DNA sequences grounded on a 2-D graphical representation. Chem. Phys. Lett. 423, 50–53.

Wolfram, S., 1984. Cellular automation as models of complexity. Nature 311, 419–424.

Wolfram, S., 2002. A New Kind of Science. Wolfram Media Inc., Champaign, IL.

Wu, Z.C., Xiao, X., Chou, K.C., 2010. 2D-MH: a web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. J. Theor. Biol. 267 (1), 29–34.

Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X., Chou, K.C., 2005a. Using cellular automata to generate image representation for biological sequences. Amino Acids 28, 29–35.

Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X., Chou, K.C., 2005b. An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation. J. Theor. Biol. 235, 555–565.

Xiao, X., Shao, S.H., Chou, K.C., 2006a. A probability cellular automaton model for hepatitis B viral infections. Biochem. Biophys. Res. Comm. 342, 605–610.

Xiao, X., Shao, S.H., Ding, Y.S., Huang, Z.D., Chou, K.C., 2006b. Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30, 49–54.

Xiao, X., Wang, P., Chou, K.C., 2008. Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image. J. Theor. Biol. 254, 691–696.

Xiao, X., Wang, P., Chou, K.C., 2009. GPCR-CA: a cellular automaton image approach for predicting G-protein-coupled receptor functional classes. J. Comput.Chem. 30, 1414–1423.

Yao, Y.H., Dai, Q., Nan, X.Y., He, P.A., Nie, Z.M., Zhou, S.P., Zhang, Y.Z., 2008. Analysis of similarity/dissimilarity of DNA sequences based on a class of 2D graphical representation. J. Comput. Chem. 29, 1632–1639.

Yao, Y.H., Nan, X.Y., Wang, T.M., 2005. Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation. Chem. Phys. Lett. 411, 248–255.

Yu, J.F., Sun, X., Wang, J.H., 2009. TN curve: a novel 3D graphical representation of DNA sequence based on trinucleotides and its applications. J. Theor. Biol. 261, 459–468.

Zhang, C.T., Chou, K.C., 1993. Graphic analysis of codon usage strategy in 1490 human proteins. J. Protein Chem. 12, 329–335.

Zhang, C.T., Chou, K.C., 1994. Analysis of codon usage in 1562 *E. Coli* protein coding sequences. J. Mol. Biol. 238, 1–8.

Zhang, C.T., 1997. A symmetrical theory of DNA sequences and its applications. J. Theor. Biol. 187, 297–306.

Zhang, C.T., Zhang, R., Ou, H.Y., 2003. The Z curve database: a graphic representation of genome sequence. Bioinformatics 19, 593–599.

Zhang, R., Zhang, C.T., 1994. Z curves, an intuitive tool for visualizing and analyzing the DNA sequences. J. Biomolecular Struct. Dynam. 11, 767–782.

Zhang, Z.J., 2009. DV-Curve: a novel intuitive tool for visualizing and analyzing DNA sequences. Bioinformatics 25, 1112–1117.

Zhao, L.P., Lv, Y.H., Li, C., Yao, M.H., Jin, X.Z., 2010. An S-curve-based approach of identifying biological sequences. Acta Biotheor. 58, 1–14.

Zhou, G.P., Deng, M.H., 1984. An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways. Biochem. J. 222, 169–176.